Of course, if the law controlling the changes of length becomes known quantitatively it will always be easy to adjust the tabulation to meet such a prerequisite.

Rule V.—As immediate consequences of the foregoing rules the results of any periodicity tabulation will be a composite aggregate of possibly several real commensurate periods, probably of small amplitude, upon which will be superposed numerous accidental errors and many uneliminated residuals, greatly obscuring the actual facts.

In the examination of the results it must constantly be recognized that sequences of wholly fortuitous numbers will always exhibit periodic features, and these, as well as all real commensurate periods, can be evaluated by the Fourier analysis or other devices. Nevertheless, only those features can be claimed as real which emerge and persist and endure in a more or less consistent fashion, regardless of some particular method of derivation. The results must be derived in as many legitimate and different ways as possible. Only those features which consistently survive and emerge from every analysis can be regarded as real periodic features in any body of data. All those features which vanish, change, and reappear incident to every legitimate change of data, method of treatment, etc., must be regarded as quite spurious, unreal, and largely the vagaries of fortuitous conditions.

The whole atmosphere can not, of course, be expected to act as a unit with respect to periodicities, and we must be prepared to find wide differences at different times and in different localities.

If it were not quite foreign to the scope and purpose of this note, it would be most interesting and instructive to show at this point the practical working of the periodicity tabulations on actual data, and the application of the rules in the interpretation of complicated results which are secured. These must, however, be reserved for another time.

551.590.2 : 551.501

FITTING STRAIGHT LINES TO DATA GREATLY SIMPLIFIED WITH APPLICATIONS TO SUN-SPOT EPOCHS

By CHARLES F. MARVIN

[Weather Bureau, Washington, March 24, 1924]

Many studies of the data of meteorology, economics, business, etc., are facilitated and definite results may be expressed by the evaluation of a straight line of best fit to the statistics involved. This is often accomplished in an approximate way by graphical methods, but in a great many cases a far more certain and accurate result can be secured by a very simple arithmetical calculation following rigorously the principle of least squares. Moreover, the computation really entails much less time and effort than that required to produce the less accurate scale drawing of the necessary chart.

The cases in which this simple method can be used arise whenever the data correspond to exactly equal and uniform intervals of time, like days, weeks, months, seasons, etc. In still other cases the observed values correspond to a series of abstract integers like 0, 1, 2, 3, etc., which represent recurrences of certain features, such, for example, as consecutive observations of the epochs or dates of the minima, or maxima of the sun-spot period. Finally, even when the original layout of the statistics does not satisfy the above simplifying condition it is often possible to make some simple adjustments of the data so that the simplifying condition is satisfied. It seems from the foregoing that there are a large number of problems in which the simple computations can be employed, and every student of statistics should be perfectly familiar with it.

The problem is to compute the best values (as defined by least squares) of the constants a and b in the general equation of the straight line,

$$y = a + bx$$

where y represents any series of observations corresponding to integral values of $x = 0, 1, 2, 3$, etc.

In order to accomplish two objects in this same note, I will ask the reader to turn his thoughts for a moment to Newcomb's method [1] of evaluating the normal epochs of sun-spot phenomena and the normal length of the period. His normal value of the sun-spot cycle 11.13 years is widely quoted and universally accepted as probably the best evaluation of this puzzling solar feature. His method must, therefore, be, as it is, a very sound one, nevertheless it seems to be little understood and almost never used, either in the analysis of modern sun-spot data not available to Newcomb or in a hundred other problems of periodicities in other statistical data.

Newcomb's method is simply that of fitting a straight line to the observations which fix the dates of the maxima, the minima, the mid-phase values rising or falling, or any other chosen characteristic of data that may be available, and since the consecutive observed values correspond to successive abstract numbers 0, 1, 2, 3, etc., representing recurrences of the same thing, the simplifying condition of the arithmetical computation is satisfied at the outset.

Both objects of this note, therefore, are accomplished by the calculation of the sun-spot data since, say 1820, to date.

Observations.—We shall use the dates given by Wolfer for simply the minima of sunspots since 1820.

TABLE 1.—*Dates of epochs of minima of sun spots by Wolfer, 1820 to 1924*

x	y	c	x	y	c
0	1,823.3	+3.3	5	1,878.9	+3.9
1	1,833.9	+2.9	6	1,889.6	+3.6
2	1,843.5	+1.5	7	1,901.7	+4.7
3	1,856.0	+3.0	8	1,913.6	+5.6
4	1,867.2	+3.2	[1] 9	1,923.9	+4.9

[1] The epoch of the present sun-spot minimum has not as yet been established accurately, but it will probably differ very little from the date indicated.

Almost every student contents himself with the faulty method of deducing the average length of the period by subtracting the first date from the last one and dividing the difference by 9, viz, $\frac{100.6}{9} = 11.18$ years. This not only presupposes that the first and last dates are exact ones, but it wholly ignores the irregular intermediate dates, and any attempted adjustment of the intermediate epochs to a normal series assigns all the irregularities to the intermediate epochs, while the first and last stand 100 per cent perfect. This is clearly wrong, because we must presuppose that each of the dates is affected by some error or irregularity and determine the amount thereof fairly by the method of analysis. This is what Newcomb's method does.

GRAPHICAL SOLUTION

Procedure.—Lay off on the Y axis of a coordinate diagram a scale of dates beginning preferably a little

[1] Astrophysical Journal, vol. 13, 1901, p. 1.

before the date of the first observed epoch of minimum to be analyzed. Lay off points on this scale locating all the observed minima.

Transfer each of these points to consecutive ordinates at uniform intervals corresponding to x values 0, 1, 2, 3, etc. (See fig. 1.) Such points in general will fall in a diagonal, nearly straight line. Fit a straight line to the points as well as possible, either by eye or by analytical methods, and the problem is solved.

Results.—The slope of the line gives the normal length of the period. This value is best found by noting on the vertical scale two intersections of the

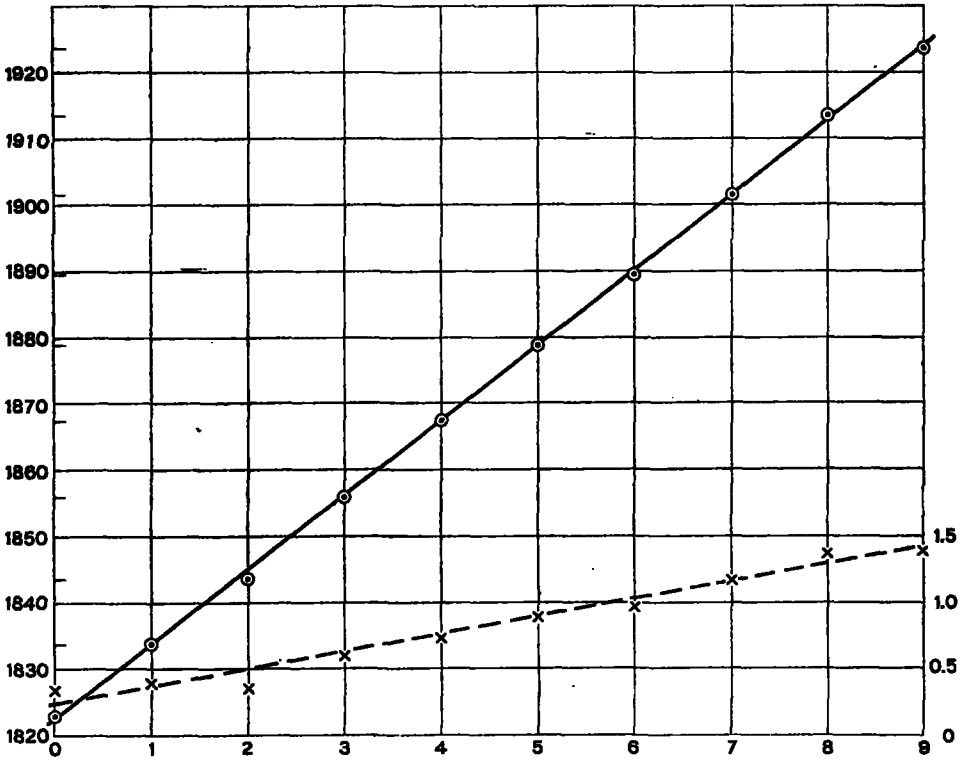

FIG. 1.—Representing the graphical solution of fitting straight lines to observations of epochs of minimum of sun spots, according to Newcomb's method

line with two widely separated ordinates. The slope is the ratio of the two sets of differences, thus:

Intersection at ordinate 9 = 1924.0
Intersection at ordinate 0 = 1822.2
Differences 9 101.8

$$\text{Average period } \frac{101.8}{9} = 11.31 \text{ years.}$$

The several intersections of the line with the ordinates 0, 1, 2, 3, etc., fix the normal or adjusted dates for the corresponding minimum epochs.

Greater accuracy is attained in the graphical solution if the vertical scale is exaggerated. Assume a uniform 10 or 11 year interval and plot on an enlarged vertical scale the excess of the observed over the assumed regular intervals. The points located by the crosses in the lower part of the diagram (scale at right) show the excesses over the uniform scale 1820, 1830, 1840, etc.

ARITHMETICAL METHOD

We are to find the best values of a and b in the equation of the straight line

$$y = 1820 + a + bx$$

Simply to avoid large numbers we assume $b = 11.0 + b'$ and our working equation now becomes

$$a + b'x = (y - 1820 + 11.x) = c$$

The quantity c is now a small number for each observation instead of the awkward large number representing the dates of the several minima. Its values are given in Table I.

Procedure.—Write down the observations (c) in two columns (I, II) and form the differences, $d = (\text{II}-\text{I})$ as indicated. Multiply these by certain weights, g, in this case 9, 7, 5, 3, and 1. Find the sum of all the c's = 36.6, also of the products $gd = 51.8$

TABLE 2.—*Calculation of b′ and a*

[$n = 10$. N (see Table 4) = 165. Weight = $n-1$, $n-3$, $n-5$, etc.]

	Observations c		Difference, II–I d	Weights g $n-1$, $n-3$, ...	Products gd
	I down	II up			
	3.3	4.9	+1.6	9	14.4
	2.9	5.6	+2.7	7	18.9
	1.5	4.7	+3.2	5	16.0
	3.0	3.6	+0.6	3	1.8
	3.2	3.9	+0.7	1	0.7
Sums	13.9	{22.7 / 13.9			
Total	Σc = 36.6				51.8

$$b' = \frac{51.8}{N = 165} = 0.314$$

Hence $b = 11.314$ which is the best normal length of the sunspot period between 1820 and 1924.

$$a = \frac{36.6}{n = 10} - \left[\frac{(n-1)}{2}\, b' = 1.413\right] = 2.247$$

Hence, adjusted date of initial epoch of minimum is, 1822.25 years, a date 1.1 years before the observed date.

Final equation, $y = 1822.25 + 11.31x$.

TABLE 3.—*Final results*

Epochs of minima observed	Epochs adjusted Calc.	Difference observed —Calc.
1823.3	1822.2	+1.1
33.9	33.6	+0.3
43.5	44.9	−1.4
56.0	56.2	−0.2
67.2	67.5	−0.3
78.9	78.8	+0.1
89.6	90.1	−0.5
1901.7	1901.4	+0.3
13.6	12.8	+0.8
23.9	24.1	−0.2

Great stress is laid by some writers upon the *variability* of the length of the sun-spot cycle, and a great deal of significance is claimed for the variations. In so far as the one hundred years of observations comprised in the above analysis are concerned *there is a very striking constancy of the period* as shown by the small residuals in Table 3, and it is difficult to see any significance to the slight fluctuations which appear.

In writing down the observations in the two columns I and II of Table 2, it is necessary that the last observation should always stand opposite the first and then the others will pair off together.

When the number of observations is odd the middle observation must, of course, stand alone at the foot of either I or II. It also occurs that the weights in this case are always *even* numbers and end at the foot of the table with 0; that is, the middle observation has no weight whatever in fixing the value of b'.

As a final comment we may suggest that it will rarely be necessary to carry out the calculations for a large number of observations, individually, but rather these may be conveniently grouped in two's, three's, five's, etc., thus reducing the large number to a series of, say, 10 or 20 values. A little judicious planning of the layout of problems suffices to bring almost any problem of this kind within the scope of the simple computations in Table 2.

For the sake of completeness we may write here the basic equations which evaluate a and b' following the calculations in Table 2. The computer needs only to follow the simple rules to which these equations lead without necessarily understanding them clearly.

$$a = \frac{\Sigma c}{n} - \frac{n-1}{2} b' \qquad (A)$$

$$b' = \frac{2\Sigma xc - (n-1)\Sigma c}{N = [2\Sigma x^2 - \frac{n}{2}(n-1)^2]} \qquad (B)$$

Now the great simplification comes in (B). Expanding the numerator leads to the combination of the observations into pairs, which can be weighted and summed as in the last column of Table 2. The proper weights are $n-1$, $n-3$, $n-5$, etc., in all cases.

Furthermore, the denominator is always a definite number depending only upon how many observations are used. This denominator, N, together with the sum of squares of the natural numbers from 1 to 25 are easily computed, once for all, and are given in Table 4. The sums of squares are really not needed in the present case, but are given as it is sometimes convenient to have them, and tables containing these values are not very numerous.

TABLE 4.—*Values of N and Σ n² for natural numbers 1 to 25*

$[N = \left[2\Sigma x^2 - \frac{n}{2}(n-1)^2\right],\ x = 0, 1, 2, 3, \text{etc.}]$

n	N	Σn^2
1	----------	1
2	1	5
3	4	14
4	10	30
5	20	55
6	35	91
7	56	140
8	84	204
9	120	285
10	165	385
11	220	506
12	286	650
13	364	819
14	455	1,015
15	560	1,240
16	680	1,496
17	816	1,785
18	969	2,109
19	1,140	2,470
20	1,330	2,870
21	1,540	3,311
22	1,771	3,795
23	2,024	4,324
24	2,300	4,900
25	2,600	5,525

551.501

ON KRICHEWSKY'S METHOD OF FITTING FREQUENCY CURVES

By EDGAR W. WOOLARD

[Weather Bureau, Washington, D. C., March 10, 1924]

A Law of Facility may be described as the approximate expression of the relative frequency with which, in the long run, different values are assumed by a quantity which is dependent on a number of variable items or elements, given certain conditions which seem to be adequately fulfilled in common experience. For example, the Law of Facility in the familar case of the ordinary errors of observation was exhaustively studied many years ago and has long been accurately represented by the so-called Gaussian curve of errors, the equation of which is well known.

In recent years the great value of being able to derive with quantitative precision the curve which shall exhibit the law of facility of a quantity under consideration has come to be realized to a greater and greater degree in an immense variety of fields of study. In any case the problem is to find from a finite number of observations, which give a more or less irregular frequency polygon or histogram, the curve which approximates most closely to the frequency curve which would result if we could have an infinite number of observations.

We now have several well-known methods of fitting curves to observed frequency distributions. The first difficulty in curve fitting is that of choosing a suitable curve from among all the possible algebraic and transcendental curves that suggest themselves; the second difficulty lies in evaluating the constants of the equation of the adopted curve. Until a comparatively recent date, the great majority of applications of the theory of frequency curves were to errors of precision measurements, which, as mentioned above, usually conform closely to the Gaussian or Normal Law. As a result, the Normal Curve became a Procrustean bed to which all possible measurements had to be made to fit; not until late in the nineteenth century did skew curves gain general recognition.[1] Again, it was for a long time taken for granted that the correct method of evaluating the

[1] See Arne Fisher. *The Mathematical Theory of Probabilities.* Vol. I, 2 ed., pp. 178–187. New York, 1922.